Coincident sequence cloning

Coincident sequence cloning

A.J.Brookes and D.J.Porteous
MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

Received March 4, 1991; Revised and Accepted April 23, 1991

ABSTRACT

We describe a novel method, Coincident Sequence Cloning (CSC), which permits the selective recovery of common sequences shared between two complex and partially coincident DNA mixtures. We evaluate this method by integrating human DNA with DNA from a mouse-human somatic cell hybrid, and we recover exclusively human DNA products which are all represented in the hybrid genome. CSC strategies should be useful in addressing many highly complex problems in genome analysis.

INTRODUCTION

Approaches to mammalian genome analysis commonly employ 'enrichment strategies' to achieve an initial significant reduction in the level of DNA complexity. One example of such a strategy would be the use of somatic cell genetics to isolate subgenomic fragments, and another the construction of cDNA libraries to access sequences having tissue restricted expression profiles. The degree of genome simplification attainable by these approaches is however limited, and therefore, the comprehensive analysis of the resulting material remains a considerable task.

A conceptual solution to this problem would be to apply different enrichment strategies sequentially to achieve a multiplicative degree of purification. Unfortunately, the majority of DNA enrichment procedures can not be combined in this way. We have therefore considered an alternative means by which the enrichment characteristics of two different strategies can be coupled, that is, to integrate directly a pair of enriched mixtures in such a way as to recover only those sequences common to both. This approach could be employed in any situation where sequences of interest are shared between complex DNA resources such as cDNA libraries, somatic cell hybrid DNA or cloned genomic DNA. To this end we have developed a method for the selective recovery of the common DNA shared between two mixtures. We propose to call this method Coincident Sequence Cloning (CSC).

In this report we evaluate CSC by integrating human DNA sequences with total DNA from a mouse-human somatic cell hybrid containing only 5% of the human genome. The products recovered in this experiment are exclusively human and all map back to the human component of the hybrid genome.

MATERIALS AND METHODS

Southern blotting and hybridisation

Southern blots were prepared as described elsewhere (1). *In vitro* labelling was by the oligo-labelling method (2). Hybridisations were performed at 68°C for 16 hours in a buffer containing 5×SSC (1×SSC comprises 0.15M NaCl, 0.015M sodium citrate pH7.4), 5×Denhart's solution, 0.1% SDS, 0.1% PPi, 10% dextran sulphate and 50μg/ml denatured sonicated salmon sperm DNA. Unless otherwise stated, post-hybridisation washes were conducted at 65°C in 0.1×SSC, 0.1% SDS.

Preparative agarose gel electrophoresis

Samples DNA's were separated on low melting temperature agarose gels in TAE buffer (0.04M Tris/acetate pH8.0, 0.001M EDTA). Regions of the gel containing DNA fragments to be recovered were excised and melted for 5 minutes at 65°C. NaCl and agarase (Calbiochem) were added to 50mM and 100u/ml respectively, and the sample placed at 37°C for 4−16 hours. The DNA was finally recovered by phenol extraction and ethanol precipitation.

Oligonucleotide sequences

The following oligonucleotides were synthesised by standard phosphoramidite chemistry and employed in the CSC procedure as described in the text:

5'capture oligo;
5' GGACGGGTCGACACGCGAGGAGCCAAGCTTGCATGCCTGCA 3'

3' capture oligo;
5' AATTCGTAATCATGGTCATAGAGCACCCGTGCTACCGGAACG 3'
485; 5' TGATTACGAATTGGTGCAGGCATGCAAG 3'

PCR Primers:

596; 5' GGACGGGTCGACACGCGAGG 3'
789; 5' CGTTCCGGTAGCACGGG 3'
790; 5' GCCAAGCTTGCATGCCTG 3'
996; 5' GCTCTATGACCATGATTACG 3'

Polymerase chain reactions

All reactions were carried out in 50μl buffer (10mM tris/HCl, 50mM KCl, 1.5mM $MgCL_2$, 0.2mM each dATP/dTTP/dCTP/dGTP, 0.01% w/v gelatin, 0.05% each Tween20 and NP40 detergents, pH8.3 at 25°C) using 2u Amplitaq enzyme and a Hybaid Intelligent Heating Block on mode 2 (plate) control. Denaturing steps were at 99°C for a duration of 2 minutes for the first cycle and 45 seconds for the remaining cycles. Extension reactions were done at 74°C for durations of 2 minutes for the first ten cycles, 2.5 minutes for the second ten cycles and 3 minutes for any further cycles. Annealing steps were of 2 minutes duration at 58°C for primers 596/789 and 52°C for primers 790/996.

CSC Procedure

EcoRI/PstI digested DNA from mixture A was cloned into M13mp18 (3). Template DNA was prepared as previously described (4,5). 5′ and 3′ capture oligos were added at a 1:1 molar ratio to 1μg of this template DNA in 40μλ 10mM tris/HCl, 1mM $MgCl_2$ pH7.5 and the mixture heated to 65°C and allowed to cool to 37°C over approximately 30 minutes. To bind any unhybridised capture oligos, oligonucleotide 485 was then added, in 1μλ H_2O, at a molar ratio to each capture oligo of 10:1. This mixture was left at 37°C for 15 minutes and then placed on ice.

15μg mixture B DNA was digested with the enzyme pair *EcoRI/PstI* and then phenol, chloroform and ether extracted and ethanol precipitated. Following resuspension in 50μλ H_2O, the sample was denatured by adding 50μλ 0.34M NaOH and placing at 37°C for 30 minutes. 100μλ of a prechilled 1:1 mixture of 0.34M HCl and 0.2M tris/HCl pH7.5 was added to neutralise the solution and the sample placed on ice. The modified DNA from mixture A was then added and the total sample ethanol precipitated. After resuspension in 18μλ H_2O both 3μλ of 4M NaCl, 50mM EDTA, 0.1M tris/HCl pH7.8 and 9μλ formamide were added and the mixture allowed to anneal by submerging overnight in a 45°C waterbath. The sample was precipitated and resuspended in 10μλ 0.5mM EDTA, 5mM tris/HCl pH7.5. A ligation was then performed at 16°C for 3 hours in 20μλ 50mM tris/HCl, 10mM $MgCl_2$, 10mM dithiothreitol, 1mM spermidine and 1mM ATP pH7.4 using 0.5u T4 DNA ligase.

The ligated DNA was passed through a 1.3% alkaline agarose gel (30mM NaOH, 2mM EDTA) from which fragments in the 0.15−0.5kb size range were recovered as gel slices to be diluted twofold in H_2O and melted at 65°C. 30 cycles of PCR were performed upon 1μλ aliquots of the liquified gel using primers 596/789. 1μλ of each of these reactions was then further amplified by 22 cycles of PCR using primers 790/996 and the products recovered by preparative electrophoresis through a 1.3% neutral agarose gel. Following *EcoRI/PstI* digestion these products were cloned, by standard procedures, into the pBluescribe plasmid vector (Statagene).

The CSC principle

A schematic representation of the CSC method applied to a pair of DNA mixtures (A and B) is illustrated in figure 1. DNA fragments from mixture A are converted into short, orientated and single stranded molecules with defined sequences at each end by first digesting with two restriction enzymes and then cloning into M13. The resulting mixture is called library A. A pair of 'capture oligos' are then annealed to library A. Mixture B is digested with the same restriction enzyme pair that were used to process mixture A and then alkali denatured. Library A and mixture B are combined and allowed to anneal in a reaction driven to completion by the components of library A. Following a ligation step, mixture B sequences are purified from those of library A by preparative alkali agarose gel electrophoresis. Coincident species are selectively recovered from this material by employing the polymerase chain reaction (PCR) with primers derived from the capture oligo sequences. PCR products are finally cloned and analysed.

RESULTS

To evaluate the CSC procedure we employed human DNA and DNA from the somatic cell hybrid 1W1 (6) (a human-mouse hybrid with an intact chromosome 11 plus part of Xpter as the sole human component) as mixtures A and B respectively, as illustrated in figure 1.

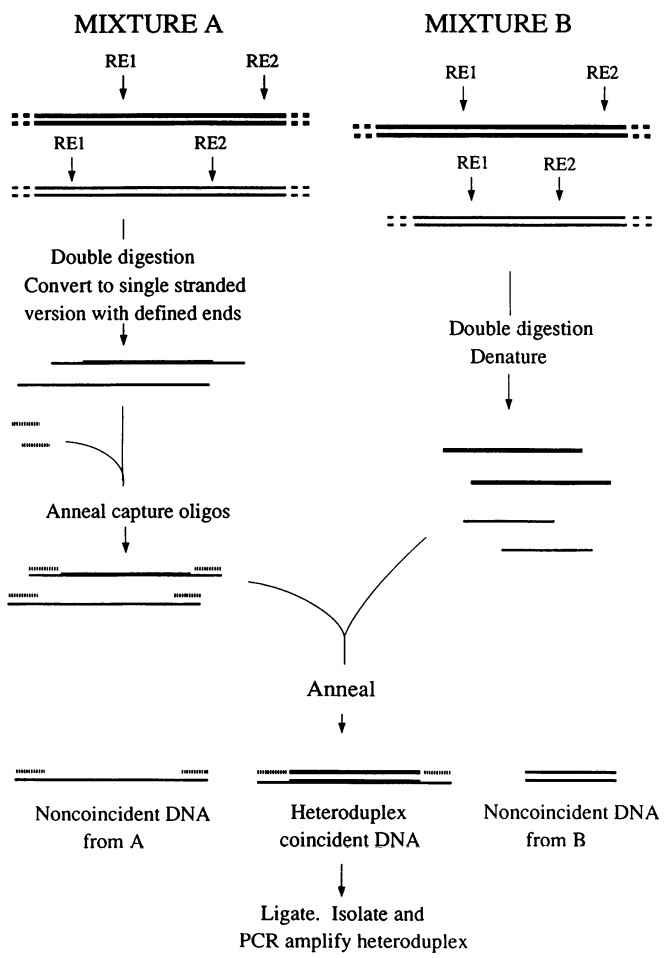


Figure 1. Schematic Representation of the CSC Procedure. An explanation of this figure is given in the text. Coincident DNA is indicated by the use of thicker lines. Dotted lines indicate capture oligos. RE1 and RE2 represent any pair of restriction enzymes.

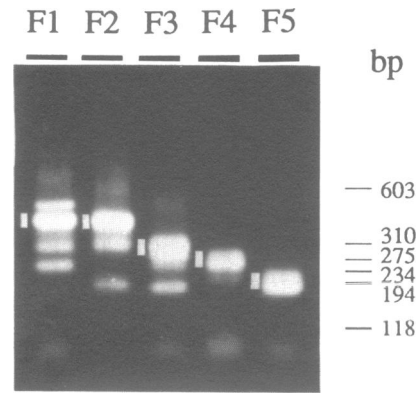


Figure 2. Preparative Agarose Gel of Coincident DNA Fractions F1-F5. Tracks marked F1-F5 show the preparative separation of coincident DNA fractions F1-F5 respectively, following PCR amplification as described in Materials and Methods. Bands recovered for the construction of product libraries are marked by solid bars.

Library A

To produce library A, total human DNA was digested with the enzyme pair *EcoRI/PstI* and fragments in the 0.15−0.5kb size range were isolated by preparative electrophoresis upon a 1.3% agarose gel for cloning into M13. 1,000 randomly selected plaques were each inoculated into 150μl minicultures. These were pooled for the preparation of single stranded DNA. 'Capture oligos' were then annealed to this library.

Integration of library A and mixture B

15μg of *EcoRI/PstI* digested 1W1 DNA was reacted with 1μg library A as detailed in the Materials and Methods. Following the ligation step coincident sequences of 0.5kb down to 0.15kb in size were isolated as a set of five 100μl fractions labelled F1−F5 respectively. These fractions were individually amplified by PCR and the main product bands (see figure 2) were recovered and cloned to give CSC product libraries F1−F5.

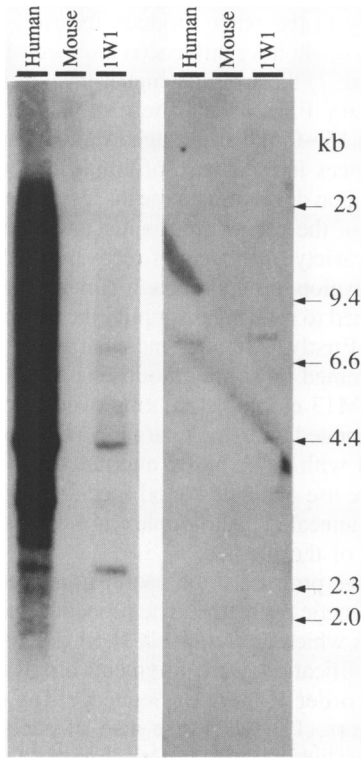
Table 1. CSC products examined and their product library of origin. Arrows indicate the relationship between product libraries and CSC products.

PRODUCT LIBRARY	NUMBER OF CLONES EXAMINED	DISTINCT PRODUCTS
1	4	1.2a
2	6	2a
3	3	3a
	14	3−1, 3−15, 3−20
4	4	4a
5	2	5a

Analysis of CSC products

To analyse the CSC products, between 2 and 6 clones were selected at random from each product library. Nitrocellulose lifts of these colonies were probed with gel purified inserts derived from miniprep plasmid DNA for each clone in order to identify duplicated products. The 2 distinct isolates so obtained from product library F3 were used as probes upon a further 22 library F3 colonies enabling 14 non-hybridising colonies to be located. From these a further 3 distinct products were similarly identified. These results are summarised in table 1.

To establish the origin of each product, all isolates were used to probe Southern blots of *EcoRI* digested human, mouse and 1W1 DNA in the absence of human or mouse competitor DNA. In all cases the probes were found to have been derived from human sequences present within 1W1, ie. from the DNA coincident between mixtures A and B. In all but one case a single hybridising band was detected with the exception showing hybridisation to a high copy number repeat element later shown to be the human L1 repeat (7,8) (data not shown). Examples of these results are shown in figure 3. By sequencing a portion of each isolate, it was found that clone 2a and all the product library F3 derived members shared a degree of homology ranging from



Product 1.2a **Product 4a**

Figure 3. Examples of Southern Analysis of CSC Products. Southern blots were prepared from *EcoRI* digests of 7μg human DNA, 20μg mouse DNA and 20μg 1W1 DNA. All CSC products were used as probes upon these blots employing high stringency post-hybridisation washes. Results obtained with a unique sequence product (4a) and an L1 repeat sequence product (2a) are shown.

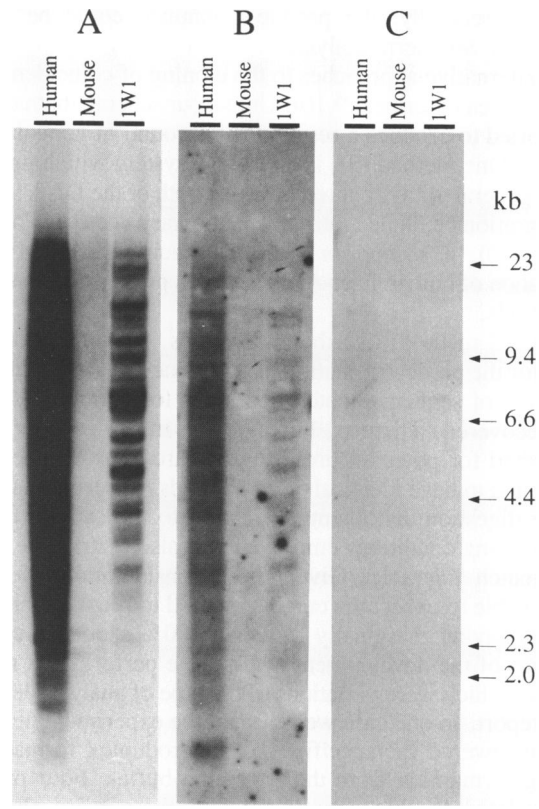


Figure 4. Southern Analysis of CSC Product 3a. Identical Southern blots were prepared from *EcoRI* digests of 7μg human DNA, 20μg mouse DNA and 20μg 1W1 DNA. CSC product 3a was used to probe these blots employing 65°C post-hybridisation washes in 0.1% SDS at various salt concentrations as shown: A; 6×SSC, B; 0.5×SSC, and C; 0.1×SSC

Table 2. Summary of results from Southern blot analysis of CSC products.

CSC PRODUCT	HUMAN +ve?	1W1 +ve?	SOUTHERN ANALYSIS (high stringency)
1.2a	+	+	Repeat
2a	+	+	Unique
3a	+	+	Unique
3−1	+	+	Unique
3−15	+	+	Unique
3−20	+	+	Unique
4a	+	+	Unique
5a	+	+	Unique

67% to 90% (data not shown). One such product was therefore used to probe genomic blots under various stringencies (see figure 4). This family of products were thus shown to be members of a previously unrecognised family of repeat elements, even though they perform as distinct single copy probes upon Southern analysis. These results are summarised in table 2.

DISCUSSION

The study described in this report was designed as a proof of principle experiment for the CSC procedure. The evaluation of any new technology requires the use of a test system which can be easily analysed. In this respect the integration of human and somatic cell hybrid DNA was an ideal system for the evaluation of CSC because firstly, these resources represent highly complex DNA and secondly, the products obtained could be readily analysed by Southern analysis.

Two alternative approaches to the cloning of coincident DNA have also been described (9,10). Unlike our scheme, both of these are reported to produce a residual background of noncoincident products. One method (9), is an *in vivo* system which promises to be of general utility, but requires that both of the DNA samples for integration be cloned into complementary vectors. The other method (10), a second *in vitro* approach, is limited to the examination of human inter-Alu sequences present within somatic cell hybrids.

Our CSC method, like the alternatives, contains no specific means for the predetermination of stringency, ie. for controlling the degree of sequence match necessary for common elements to be recovered. There is, however, an innate requirement in our method for potential products to share approximate length identity and to have identical 5′ and 3′ ends (to permit restriction enzyme digestion and eventual ligation to the capture oligos). The annealing conditions employed will also restrict the degree of mismatch tolerable. Given these considerations, it is not unreasonable to expect the current method to recover coincident elements having marginally less than 100% sequence identity.

As part of the development of CSC we performed a number of studies which were variations of the type of analysis described in this report. In one case we repeated the experiment described here but lowered the specificity of heteroduplex formation by omitting formamide from the annealing buffer. Four products were isolated in this experiment and these were shown to represent a subset of the products obtained when including formamide. Notably, a significantly greater degree of amplification was required to produce these four products. One possible reason for this could be that the desired formation of matched heteroduplex was proceeding in competition with lower specificity sequence pairing so causing a smaller amount of material to become available for ligation to the capture oligos. In a second experiment we utilised as library A, 300 human DNA fragments of up to 2kb in size. Using longer extension times in the PCR reactions we derived five products ranging in length up to 1.2kb. This might indicate an intrinsic upper size limit to the procedure which could be due to, for example, the genomic distribution of short interspersed repeats.

The 8 products examined in this study are all coincident. In light of the above considerations, those products of repetitive nature could potentially have been 'captured' by highly related, though nonidentical sequences in library A. Significantly, no products were derived from the vast excess of mouse DNA (98%) present in the hybrid genome. Noncoincident sequences from library A were also not recovered since all products mapped back to 1W1. This total absence of noncoincident products is consistent with all our other studies.

The last stage of the CSC procedure involves bacterial transformation with cloned PCR products. The size of the product libraries one could make is therefore virtually unlimited. The number of distinct products present within these libraries will however have some limitation (for example, 50 of the 1,000 human fragments of library A should be present in 1W1 since this comprises 5% of the human genome). The number of distinct products we examined was not exhaustive, and, in the case of product library F3, this was proven by expanding the initial 2 isolates to a group of 5 by simply examining more colonies. Additional discrete F3 products have since been detected though not yet thoroughly examined. It is reasonable to assume that the same will apply to the other product libraries.

Three factors might have influenced the spectrum of coincident products isolated; 1) the arbitrary higher level of analysis applied to product library F3, 2) the pattern of *EcoRI/PstI* fragments present in the 0.15−0.5kb size range examined and 3) variations between sequences in their ease of amplification by PCR. We anticipate that upon developing a means by which we can control the stringency of the procedure, as discussed below, then both the range and variety of sequences recovered will be increased.

To further develop our approach, a number of methodological changes, designed to refine and simplify the procedure, are under investigation. Firstly, we are increasing the size range of molecules examined by simply modifying the PCR parameters. Secondly, the M13 cloning step, employed to convert mixture A into the orientated and single stranded DNA of library A, is being replaced with PCR based alternatives. Thirdly, we are investigating the use of single-strand specific modification and/or degradation of annealed heteroduplex as a means for controlling the stringency of the method.

CSC strategies promise to be useful in addressing a number of molecular genetic problems. The procedure necessarily gives rise to products which are relatively short due to its dependence upon PCR amplification. Such fragments are of an ideal size for sequencing in order to derive 'Sequence Tagged Site' (STS) genomic markers (11). They are also of adequate length for immediate use as hybridisation probes. For these reasons the CSC scheme is intrinsically applicable to the analysis of genomic DNA. For example, the procedure could be applied to the examination of a panel of human clones derived from a somatic cell hybrid to obtain a subset of these which are also present in a second hybrid sharing only a small degree of overlap with the first. It might also be possible to directly integrate DNA between species to access conserved sequences of functional significance. Given

an adequate means for stringency control, it might even be possible to directly localise disease genes by integrating the DNA of unrelated, affected individuals in order to identify sequences in linkage disequilibrium with the mutant locus.

Aside from genomic DNA studies we believe that CSC could be of great utility in allowing one to screen very large stretches of genomic DNA (whether in its native state or cloned in a vector system) for unknown coding sequences which exhibit a tissue restricted expression profile by using an appropriate cDNA resource as mixture B. The structural differences in sequence between genomic DNA and cDNA due to the presence of introns in the former could easily be overcome by employing frequent cutting restriction enzymes to examine small fragments and then focusing on the untranslated regions within cDNA which are generally coded within large exons. Development of the CSC procedure to address this problem is now underway.

ACKNOWLEDGEMENTS

The authors would like to thank Doreen Chambers for providing the oligonucleotides employed in this work and Barbara Stevenson for her excellent technical assistance.

REFERENCES

1. Southern, E. (1975) *J. Mol. Biol.* 98, 503–508.
2. Feinberg, A. P. and Vogelstein, B. (1984) *Anal. Biochem.* 137, 266–267.
3. Norrander, J., Kempe, T. and Messing, J. (1983) *Gene* 26, 101–106.
4. Denti, L., Cesareni, G. and Cortese, R. (1983) *Nucleic Acids Res.* 11, 1645–1655.
5. Scheier, P. H. and Cortese, R. (1979) *J. Mol. Biol.* 129, 169–172.
6. Porteous, D.J., Morten, J.E.N., Cranston, G., Fletcher, J.M., Mitchell, A., vanHeyningen, V., Fantes, J.A., Boyd, P.A. and Hastie, N.D. (1986) *Mol. Cell. Biol.* 6, 2223–2232.
7. Musich, P. R., Brown, F. L. and Maio, J. J. (1980) *Chromosoma* 80, 331–348.
8. Maio, J. J., Brown, F. L. McKenna, W. G. and Musich, P. R. (1981) *Chromosoma* 83, 127–144.
9. Kurnit, D. M. and Seed, B. (1990) *Proc. Natl. Acad. Sci. USA* 87, 3166–3169.
10. Aslanidis, C. and deJong, P. (1991) *Proc. Natl. Acad. Sci. USA* in press.
11. Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) *Science* 245, 1434–1435.